Research Statement

Ryan Giordano
rgiordan@mit.edu

November 29, 2020

As statistical models grow in size and complexity to service modern scientific datasets and questions, key data science tasks become both more important and more computationally onerous.

- Cross validation (CV) is used ubiquitously in machine learning to evaluate model predictive performance and tune hyperparameters, but requires fitting a model multiple times with different data subsets left out.
- Prior specification is a necessary step in Bayesian statistics, a statistical paradigm that provides interpretable, coherent uncertainty quantification for scientific questions. But Bayesian inference can be sensitive to the prior specification, and evaluating the model for multiple plausible prior choices can be computationally prohibitive.
- Uncertainty propagation, i.e., allowing the inferential uncertainty in one modeling quantity to inform the inferential uncertainty in another, is a key advantage of Bayesian statistics. However, the classical tool for Bayesian estimation, Markov Chain Monte Carlo (MCMC), requires evaluating a statistical model many times, and so can be prohibitively computationally expensive.

My research uses *sensitivity analysis* to provide fast, accurate approximations to these fundamental tasks of data science, providing modern theory and practical software to provide orders-of-magnitude speedup over classical methods — all with modern theoretical guarantees.

Consider uncertainty propagation in the construction of astronomical catalogues from telescopic image data. State-of-the art techniques employ a fast Bayesian posterior approximation known as "mean-field variational Bayes" [Regier et al., 2019] which is known to fail to propagate uncertainty and underestimate standard errors [Turner and Sahani, 2011]; full posterior inference using classical Markov Chain Monte Carlo (MCMC) techniques being too computationally demanding for such a large-scale problem.

The problem of prior specification is fundamental to Bayesian statistics. The Bayesian approach provides powerful tools for quanitifying uncertainty of complex quantities, such as the number of distinct clusters in a human

genomics dataset [Huang et al., 2011, Raj et al., 2014]. However, in complicated Bayesian models, it can be hard to know *a priori* how sensitive one's conclusions are to one's prior speicification. When even a single posterior approximation is extremely expensive, running for a large number of alternative prior specifications is prohibitive.

The technique of "cross validation" is a ubiquitous machine learning tool for evaluating predictive performance and tuning hyperparameters, but it requires evaluating one's model at many slightly different datasets, each with different datapoints left out [Barnard, 1974, Friedman et al., 2001]. For example, in order to select a smoothing parameter in a preprocessing step for clustering the time series of gene expression levels in mice [Shoemaker et al., 2015], one would typically choose an amount of smoothing which gives the best predictive performance according to a large number of CV fits.

These three problems of uncertainty propagation, prior sensitivity, and cross validation may seem superfically distinct, other than that all are fundamental to the practice of data science. But they share the common property that they are computationally demanding due to requiring the evaluation or estimation of a statistical model multiple times: once for each draw of an MCMC chain, once for each prior specification, or once for each cross validation sample. In turn, I show in my work that this commonality implies that they are all amendable to *sensitivity analysis*, in which the evaluation of a model at alternative inputs is approximated by a Taylor series, evaluated at a single initial model fit.

My research employs sensitivity analysis to provide accurate approximations to uncertainty propagation, prior sensitivity, cross validation and other fundamental data science problems, typically providing good accuracy and orders-of-magnitude speedups over classical approaches. A recurrent theme of my work is adapting venerable classical theoretical tools build on Taylor Series expansions [Reeds, 1976, Gustafson, 1996, Opper and Saad, 2001] to *modern computing environments* equipped with scalable, general purpose automatic differentiation software [Baydin et al., 2017, Carpenter et al., 2015].

Data sensitivity: cross validation and frequentist variance

Accuracy bounds for approximate cross validation. To perform leave-one-out CV (LOO-CV), one re-runs an estimation procedure with each datapoint left out. In full, LOO-CV requires as many re-runs procedures as there are datapoints, and each re-run is expected to be quite close to the original fit. Rather than re-running exactly, one can use a Taylor series to approximate the effect of removing a single data point; since the dataset with one point left out is, in some sense, "close" to the original dataset, the Taylor series can be expected to perform well.

In Giordano et al. [2019b], we synthesized the classical statistics and machine learning treatment of this idea [Jaeckel, 1972, Shao and Tu, 2012, Rad and Maleki, 2018, Koh and Liang, 2017], adapting the classical theory to the demands of modern machine learning. In particular, we provide finite-sample accuracy bounds on the accuracy of the Taylor series (from which previous asymptotic

results followed as a corollary), and remove the classical assumption of bounded gradients, a condition that almost never obtains in practice, but which was required by almost all previous theoretical work. We demonstrated the accuracy of the technique on an unsupervised clustering problem from genomics [Shoemaker et al., 2015].

Sensitivity to removal of a small fraction of the data. Classical frequentist standard errors estimate the variability in an estimator that would result from the rarefied thought experiment of re-sampling datasets from the same distribution that gave rise to the observed data. In the social sciences, this rarefied experiment rarely closely corresponds to reality, and one might be concerned if substantive conclusions could be overturned by other minor perturbations to the data.

In Giordano et al. [2020], we provide an easily-computed approximation to quantify the effect of ablating a small proportion of a dataset, with open-source software and finite-sample accuracy bounds for ordinary least squares and instrumental variables regression. We find that problems with small signal-to-noise ratio but large datasets will be particularly non-robust to the removal of a small proportion of the data. Such a situation that obtains commonly in econometrics, and we find that the sign and statistical significance of estimated effects in a number of large, prominent econometric studies can be overturned by dropping only a small number of datapoints [Angelucci and De Giorgi, 2009, Finkelstein et al., 2012, Meager, 2019].

Frequentist variability of Bayesian posteriors. Bayesian statistics provides powerful tools for coherently treating uncertainty in complex problems, though, when the model is misspecified, the estimated posterior uncertainty may not be meaningful. In principle, however, one might always compute the frequentist sampling variability of a Bayesian posterior quantity, and such a quantity always remains meaningful, even if conceptually distinct from a posterior uncertainty [Waddell et al., 2002, Kleijn and van der Vaart, 2006]. However, standard tools for evaluating frequentist uncertainty, such as the bootstrap [Huggins and Miller, 2019], are extremely computationally intensive, as they typically require re-running an MCMC procedure hundreds of times.

In a work in progress [Giordano and Broderick, 2020], we derive the Bayesian infinitesimal jackknife (IJ), which we prove can be used to consistently estimate the frequentist variability of Bayesian posterior means without bootstrapping or computing a maximum a-posteriori (MAP) estimate. Our work synthesizes results from Bayesian robustness and frequentist von Mises expansions and extends the Bayesian central limit theorem to the expectation of data-dependent functions [Jaeckel, 1972, Shao and Tu, 2012, Giordano et al., 2019b, Gustafson, 2000, Giordano et al., 2018a, Lehmann and Casella, 2006, Kass et al., 1990]. We demonstrate the accuracy of the Bayesian IJ on datasets from election modeling [Gelman and Heidemanns, 2020], ecology [Kéry and Schaub, 2011], and most of the models from [Gelman and Hill, 2006, Stan Team, 2017], showing that

the Bayesian IJ can reproduce the bootstrap covariance estimates in orders of magnitude less compute time.

Sensitivity for Bayesian analysis

Propagation of uncertainty in scalable Bayesian inference One popular technique to scale Bayesian inference to massive problems is mean field variational Bayes (MFVB) [Wainwright and Jordan, 2008, Blei et al., 2017, Regier et al., 2019]. However, MFVB provides notoriously innacurate posterior uncertainty estimates, even in situations when it estimates the posterior means accurately [Turner and Sahani, 2011]. In Giordano et al. [2018a], we develop a method to recover accurate posterior uncertainties from MFVB approximations without needing to fit a more complex model or run MCMC. Computing the LRVB covariance requires solving a linear system, which in scientific applications is often sparse and can be solved using iterative techniques such as conjugate gradient [Nocedal and Wright, 2006, Chapter 5].

We compare LRVB covariances to MCMC on a large number of real-world datasets, including logistic regression on an internet advertising dataset [Criteo Labs, 2014], the Cormack-Jolly-Seber model from ecology [Kéry and Schaub, 2011], and hierarchical generalized linear models from the social sciences [Gelman and Hill, 2006], demonstrating accurate posterior covariances computed over an order of magnitude faster than MCMC.

Hyperparameter sensitivity for MCMC. MCMC is arguably the most commonly used computational tool to estimate Bayesian posteriors, and modern black-box MCMC tools such as `Stan` [Stan Development Team, 2020, Carpenter et al., 2017]. However, MCMC is typically still time-consuming, and systematically exploring alternative prior parameterizations by re-running MCMC would be computationally prohibitive for all but the simplest models. A classical result from Bayesian robustness states that the sensitivity of a posterior expectation is given by a particular posterior covariance [Gustafson, 1996, Basu et al., 1996], though the result has not been widely used, arguably due in part to the lack of an automatic implementation. In my software package, Giordano [2020], I take advantage of the automatic differentiation capacities of `Stan` to provide automatic hyperparameter sensitivity for generic Stan models. In examples in the package `git` repository, I demonstrate the efficacy of the package in detecting excess prior sensitivity, particularly in a social sciences model taken from Gelman and Hill [2006, Chapter 13.5].

Bayesian nonparametrics. A commonly question in unsupervised clustering is how many distinct clusters are present in a dataset. Discrete Bayesian nonparametrics (BNP) allows the answer to be inferred using Bayesian inference, but one must specify a prior on how distinct clusters are generated [Ghosh and Ramamoorthi, 2003, Gershman and Blei, 2012]. A particularly common modeling choice is the stick-breaking representation of a Dirichlet process prior

[Sethuraman, 1994], a mathematical abstraction which is arguably better justified by its computational convenience than its realism.

In Giordano et al. [2018b], we fit a BNP model with variational Bayes [Blei and Jordan, 2006] using the standard, computationally convenient stick-breaking prior, but then use sensitivity analysis to allow the user to explore alternative functional forms an order of magnitude faster than would be possible with refitting. In work currently in progress, we apply our method to a human genome dataset in phylogenetics taken from [Huang et al., 2011], and find that our method accurately discovers meaningful prior sensitivity in a BNP version of the model `fastSTRUCTURE` [Raj et al., 2014].

Selected Future work

My research is driven by the needs of my scientific collaborators, and so my future work will be determined to a large part by my colleagues. Here, I will discuss a few directions that I find promising and interesting, and which I believe could be applicable to a diverse set of problems.

The higher-order infinitesimal jackknife for the bootstrap. In the preprint Giordano et al. [2019a], we extend Giordano et al. [2019b] to higher-order Taylor series approximations, provding a family of estimators which we collectively call the higher-order infintiesimal jackknife (HOIJ). In addition to providing higher-quality approximations to CV and extending our results to k-fold CV, the higher-order approach promises to provide a scalable alternative to the bootstrap, a procedure that estimates frequentist variability by repeatedly re-evaluating a model at datasets drawn with replacement from the observed data. The bootstrap is known to enjoy higher-order accuracy in certain circumstances Hall [2013], and the HOIJ can approach the bootstrap at a rate faster than the bootstrap approaches the truth. The HOIJ thus promises to make bootstrap inference available to models which are differentiable but too expensive to re-evaluate (e.g. simulation-based models [Gourieroux and Monfort, 1993]), but also to allow efficient bootstrap-after-bootstrap procedures which that are currently out of reach for all but the simplest statsitics [Efron and Tibshirani, 1994].

Partitioned Bayesian inference. The ideas of [Giordano et al., 2018a] can be naturally extended to approximately propagate uncertainty amongst separately estimated components of an inference problem. For example, astronomical catalogues are customarily produced with MFVB-like algorithms [Lang et al., 2016, Regier et al., 2019], which take inputs such as the sky background and optical point spread function as fixed inputs, though these quantities are themselves inferred with uncertainty. Viewing all the separate inference procedures as a sequential quasi-MFVB objective, one could directly apply the techniques of LRVB to propagate the uncertainty from the modeling inputs to the astronomical catalogue's uncertainty. Doing so would require the approximate solution

of a very large, but very sparse, linear system, which is itself an interesting computational challenge.

References

Angelucci, M. and De Giorgi, G. (2009). Indirect effects of an aid program: how do cash transfers affect ineligibles' consumption? *American Economic Review*, 99(1):486–508.

Barnard, G. (1974). Discussion of "Cross-validatory choice and assessment of statistical predictions". *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):133–135.

Basu, S., Jammalamadaka, S. R., and Liu, W. (1996). Local posterior robustness with parametric priors: Maximum and average sensitivity. In *Maximum Entropy and Bayesian Methods*, pages 97–106. Springer.

Baydin, A., Pearlmutter, B., Radul, A., and Siskind, J. (2017). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–153.

Blei, D. M. and Jordan, M. I. (2006). Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of statistical software*, 76(1).

Carpenter, B., Hoffman, M., Brubaker, M., Lee, D., Li, P., and Betancourt, M. (2015). The stan math library: Reverse-mode automatic differentiation in c++. *arXiv preprint arXiv:1509.07164*.

Criteo Labs (2014). Criteo conversion logs dataset. Downloaded on July 27th, 2017.

Efron, B. and Tibshirani, R. (1994). *An Introduction to the Bootstrap*. CRC press.

Finkelstein, A., Taubman, S., Wright, B., Bernstein, M., Gruber, J., Newhouse, J., Allen, H., Baicker, K., and Oregon Health Study Group (2012). The oregon health insurance experiment: evidence from the first year. *The Quarterly journal of economics*, 127(3):1057–1106.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gelman, A. and Heidemanns, M. (2020). The Economist: Forecasting the us elections. Data and model accessed Oct., 2020.

Gelman, A. and Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press.

Gershman, S. and Blei, D. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian nonparametrics*. Springer Science & Business Media.

Giordano, R. (2020). StanSensitivity: automated hyperparameter sensitivity for Stan models.

Giordano, R. and Broderick, T. (2020). *The Bayesian Infinitesimal Jackknife for Variance.*

Giordano, R., Broderick, T., and Jordan, M. (2018a). Covariances, robustness and variational Bayes. *The Journal of Machine Learning Research*, 19(1):1981–2029.

Giordano, R., Jordan, M. I., and Broderick, T. (2019a). A higher-order Swiss army infinitesimal jackknife. *arXiv preprint arXiv:1907.12116*.

Giordano, R., Liu, R., Jordan, M. I., and Broderick, T. (2018b). Evaluating sensitivity to the stick breaking prior in Bayesian nonparametrics. *arXiv preprint arXiv:1810.06587*.

Giordano, R., Meager, R., and Broderick, T. (2020). *An Automatic Finite-Sample Robustness Metric: Can Dropping a Little Data Change Conclusions?*

Giordano, R., Stephenson, W., Liu, R., Jordan, M. I., and Broderick, T. (2019b). A Swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR.

Gourieroux, C. and Monfort, A. (1993). Simulation-based inference: A survey with special reference to panel data models. *Journal of Econometrics*, 59(1-2):5–33.

Gustafson, P. (1996). Local sensitivity of posterior expectations. *The Annals of Statistics*, 24(1):174–195.

Gustafson, P. (2000). Local robustness in Bayesian analysis. In Insua, D. R. and Ruggeri, F., editors, *Robust Bayesian Analysis*, volume 152. Springer Science & Business Media.

Hall, P. (2013). *The Bootstrap and Edgeworth Expansion*. Springer Science & Business Media.

Huang, L., Jakobsson, M., Pemberton, T., Ibrahim, M., Nyambo, T., Omar, S., Pritchard, J., Tishkoff, S., and Rosenberg, N. (2011). Haplotype variation and genotype imputation in African populations. *Genetic epidemiology*, 35(8):766–780.

Huggins, J. and Miller, J. (2019). Using bagged posteriors for robust inference and model criticism. *arXiv preprint arXiv:1912.07104*.

Jaeckel, L. (1972). The infinitesimal jackknife, memorandum. Technical report, MM 72-1215-11, Bell Lab. Murray Hill, NJ.

Kass, R., Tierney, L., and Kadane, J. (1990). The validity of posterior expansions based on Laplace's method. *Bayesian and Likelihood Methods in Statistics and Econometrics*.

Kéry, M. and Schaub, M. (2011). *Bayesian population analysis using WinBUGS: a hierarchical perspective*. Academic Press.

Kleijn, B. and van der Vaart, A. (2006). Misspecification in infinite-dimensional Bayesian statistics. *The Annals of Statistics*, 34(2):837–877.

Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *International Conference on Machine Learning (ICML)*.

Lang, D., Hogg, D., and Mykytyn, D. (2016). The Tractor: Probabilistic astronomical source detection and measurement. *ascl*, pages ascl–1604.

Lehmann, E. and Casella, G. (2006). *Theory of point estimation*. Springer Science & Business Media.

Meager, R. (2019). Understanding the average impact of microcredit expansions: A Bayesian hierarchical analysis of seven randomized experiments. *American Economic Journal: Applied Economics*, 11(1):57–91.

Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

Opper, M. and Saad, D. (2001). *Advanced Mean Field Methods: Theory and Practice*. MIT press.

Rad, K. and Maleki, A. (2018). A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv Preprint*.

Raj, A., Stephens, M., and Pritchard, J. (2014). fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573–589.

Reeds, J. (1976). *Ill (1976). On the definition of von Mises functionals*. PhD thesis, Ph. D. Thesis, Statistics, Harvard University.

Regier, J., Fischer, K., Pamnany, K., Noack, A., Revels, J., Lam, M., Howard, S., Giordano, R., Schlegel, D., and McAuliffe, J. (2019). Cataloging the visible universe through Bayesian inference in Julia at petascale. *Journal of Parallel and Distributed Computing*, 127:89–104.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica*, pages 639–650.

Shao, J. and Tu, D. (2012). *The Jackknife and Bootstrap*. Springer Series in Statistics.

Shoemaker, J. E., Fukuyama, S., Eisfeld, A. J., Zhao, D., Kawakami, E., Sakabe, S., Maemura, T., Gorai, T., Katsura, H., Muramoto, Y., Watanabe, S., Watanabe, T., Fuji, K., Matsuoka, Y., Kitano, H., and Kawaoka, Y. (2015). An ultrasensitive mechanism regulates influenza virus-induced inflammation. *PLoS Pathogens*, 11(6):1–25.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.2.

Stan Team (2017). Stan example models wiki. Referenced on June 5th, 2020.

Turner, R. E. and Sahani, M. (2011). Two problems with variational expectation maximisation for time-series models. In Barber, D., Cemgil, A. T., and Chiappa, S., editors, *Bayesian Time Series Models*.

Waddell, P., Kishino, H., and Ota, R. (2002). Very fast algorithms for evaluating the stability of ml and Bayesian phylogenetic trees from sequence data. *Genome Informatics*, 13:82–92.

Wainwright, M. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Machine Learning*, 1(1-2):1–305.